Naval Aerospace Medical Research Laboratory

NAMRL-1400

THE EFFECT OF PRESENTATION MEDIUM ON PILOT SELECTION TEST BATTERY SCORES

S. Biggerstaff, D. Blower, C. Portman, and
A. D. Chapman

19980807 060

Naval Aerospace Medical Research Laboratory
51 Hovey Road
Pensacola, Florida 32508-1046

Reviewed and approved 4 Mar 98

L. H. FRANK, CAPT, MSC USN
Commanding Officer

This research was sponsored by the Naval Medical Research and Development Command under work unit 62233N MM33P30-01 7602.

Volunteer subjects were recruited, evaluated, and employed in accordance with the procedures specified in the Department of Defense Directive 3216.2 and Secretary of the Navy Instruction 3900.39 series. These instructions are based upon voluntary informed consent and meet or exceed the provisions of prevailing national and international guidelines

NAVAL AEROSPACE MEDICAL RESEARCH LABORATORY
51 HOVEY ROAD, PENSACOLA, FL 32508-1046

NAMRL-1400

THE EFFECT OF PRESENTATION MEDIUM ON PILOT SELECTION TEST BATTERY SCORES

S. Biggerstaff, D. Blower, C. Portman, and A. D. Chapman

ABSTRACT

The use of computer-based testing systems for selection and classification has gained popularity in both the civilian and military world. Even so, there are several issues needed to be resolved related to the transition of paper-and-pencil tests to computerized testing. The American Psychological Association has set guidelines to be followed to ensure both qualitative and quantitative equivalence of new test formats prior to their use in applied settings. Attention must be paid to the psychometric properties of the test design and the human-machine interface to produce a reliable and valid test. Other peripheral issues such as visual display qualities and input devices must also be carefully considered. The purpose of the study was to conduct an initial evaluation of the equivalence between the current paper-and-pencil version of the U.S. Navy's Aviation Selection Test Battery (ASTB) and a Windows-based version of the ASTB. Eighty-two U.S. Navy aviation flight school candidates at the Naval Aviation Schools Command volunteered to participate. Participants were given either the current paper-and-pencil version or the computer-based version of the ASTB or an alternate test developed to measure similar psychological attributes. We found no main effects due to medium (paper-and-pencil vs. computer-based), although performance on the alternate test was significantly worse than on the ASTB. Predictive validation of the computerized ASTB will not be completed until participant training performance data are collected.

Acknowledgments

Special thanks go out to School's Command at NAS Pensacola. Without their support, this research would have not been possible. The authors would also like to thank H. Williams, L. Temme, and K. Mayer for their editorial contributions.

INTRODUCTION

Previous research conducted in the area of computer-based testing has addressed several issues related to the transition of a test battery from a paper-and-pencil to a computerized format. Some of the advantages of computerized testing include reductions in testing time, the potential for immediate performance feedback, the ability to measure response latency, as well as the ability to collect additional information on test-taking patterns such as which items are skipped and answered last during a test session (Wise & Plake, 1989). In contrast to paper-and-pencil tests, computer-based testing allows for individualized assessment, increased capabilities in utilizing information, enhanced economic value, manipulation of measurement databases, and improved diagnostic testing (Ward, 1984).

According to the American Psychological Association (APA) Guidelines on Computer-Based Tests and Interpretations (1986), the equivalence of scores from computerized versions should be well established and documented, and comparison studies of computerized and conventional testing should be reported to establish the relative reliability of computerized assessment (Green, 1991; Van de Vijver & Harsveld, 1994). To ensure that performance on a test battery is not differentially affected by the medium of presentation, data should be collected to demonstrate equivalence. Equivalence has been defined in the literature as either qualitative (or structural) or quantitative. Qualitative equivalence refers to the notion that both forms of the test are assessing the same psychological construct, and it is typically determined by linear structural models or item factor analysis. Quantitative equivalence refers to the comparability of results between the two modes of presentation. Numerical score distributions obtained from equivalent tests must be identical or should be made identical through score transformations.

Overall, previous research has established the crucial importance of demonstrating both qualitative and quantitative equivalence between modes of administration to ensure test validity and reliability. To demonstrate equivalence, technical aspects of the computer administration must be addressed by the test creators. Depending on the perceptual and cognitive dynamics of the new test item presentation, there may be decreases in test performance and lowered validity. Mead and Drasgow (1993) found that the effect of the medium of presentation depends on the test type. Tests that contain a large number of easy items to be completed in a short time are called speeded tests. They often discriminate between individuals in the number of items answered per unit time and not in the number of items correct. Power tests are tests which have a small number of difficult items, and the amount of time allotted does not have a strong effect on performance. Mead and Drasgow (1993) found that performance on speeded tests was affected by the test medium, whereas the mode of test administration had little effect on reliability and validity in power tests.

Careful attention must be given to the psychometric aspects of the test design, as well as the human-machine interface, to produce a reliable and valid test. Features such as test item omission, backtracking and item review, or delaying responses, which are often not available in a computerized test, can impact performance. Of course, computer graphics quality and visual display format of test questions can also affect equivalence. Lastly, the input device utilized may affect the rate and ease of response. For example, pressing a key or using a mouse may actually be easier and quicker than marking an answer sheet (Federico, 1992). Each of these factors must be examined to assess the potential implications on test performance. Despite these concerns, a number of researchers have demonstrated good equivalence between computer and paper-and-pencil tests within both the military (Moreno, Wetzel, McBride, & Weiss, 1984; Kiely, Zara & Weiss, 1986; Federico, 1992) and civilian testing environments (Harrell, Honaker, Hetu & Oberwager, 1987; Vansickle, Kimmel, & Kapes, 1989; Van de Vijver & Harsveld, 1994; King & Miles, 1995).

The use of computer-based test delivery systems has gained gradual support in the military ever since computer hardware and software became more readily available at a reasonable cost in the early 1970s. The Department of Defense (DoD) research community recognized the potential of computer-based selection systems in the early 1980s, and the U. S. Navy and Air Force initiated the development of such systems independently. These parallel efforts resulted in the development of the Air Force Basic Attributes Test (BAT) and the Naval Aerospace Medical Research Laboratory (NAMRL) computer-based performance test batteries. The Air Force committed to full-scale implementation of the BAT in 1989, with the system becoming fully operational in 1991. The Navy's test battery was never made operational within the aviation community, but it has been transitioned subsequently to other communities. In the late 1980s, a tri-service Aircrew Selection Joint Planning Group stated that there was a need for a joint world-wide network for computerized testing and a common aircrew performance database. The Navy's role in this effort was to include the automation of their existing computer-based performance tests, as well as the computerization of the Aviation Selection Test Battery (ASTB). The U.S. Air Force research plans also included automation of their paper-and-pencil test, the Air Force Officer Qualifying Test (AFOQT). The project goals for the computerization of the ASTB were to create a series of tests parallel to the existing paper-and-pencil components of the ASTB. The resulting computer-based ASTB must demonstrate equivalency to the existing paper-and-pencil ASTB to meet the operational needs for a valid and equivalent form. The project involved both the development of the test sections, determination of appropriate interface features of the program, and the eventual creation of two valid forms of the computerized ASTB. The ultimate goal of all of these efforts was a single DoD-wide computerized selection test battery. The two computer-based ASTB forms were completed in Nov 1995.

The purpose of this study was to conduct an initial evaluation of the equivalence of the Windows-based Aviation Selection Test Battery (ASTB) with the existing operational paper-and-pencil version of the test. This study is part of a larger project to investigate the underlying structure of the Navy's ASTB and an in-house alternate test battery.

METHODS

SUBJECTS

Eighty-two (3 females, 79 males) U.S. Navy aviation flight school candidates at the Naval Aviation Schools Command (NASC) in Pensacola, Florida, volunteered to participate in this study. All subjects had been previously selected for pilot training in the U.S. Navy. The average age was 23.82 (*SD* = 2.01). For handedness, 76 were right handed, 3 left handed and 3 ambidextrous. Sixty-seven claimed to be designated as pilot and 15 as naval flight officers. When asked about previous flight hours, 58 had none, 13 had 1-100 h, 9 had 101-1000 h, and 2 had more than 1000 h. Two of the requirements for pilot training selection are (1) at least a bachelors degree from an accredited university and (2) scoring at or above the minimum cut-off score on the Navy's ASTB. Therefore, all participants had at least one previous experience with the ASTB before participating in this experiment. All subjects were informed as to the purpose of the study and of their freedom to withdraw at any time without prejudice to their military career.

MATERIALS

The Aviation Selection Test Battery (ASTB)

The ASTB is a multiple-choice test that was developed and validated jointly by the Naval Aerospace and Operational Medical Institute and the Educational Testing Services to predict initial ground school and primary flight training performance in the Navy, Marine Corps,and Coast Guard pilot and naval flight officer (NFO) curricula. Ground school includes coursework in basic engine properties, aerodynamics and navigation at NASC, as well as more advanced academic work on specific aircraft systems during the earliest phase of their flight training. For student naval aviators (SNAs), primary training lasts approximately 4 months and includes the above-mentioned systems coursework, flight simulator sessions, and actual flight time in the T-34C Mentor. Primary training is done at squadrons at Naval Air Station (NAS) Whiting Field, Forida., and at NAS Corpus Christi, Texas. Subsequent to their completion of primary training, students are placed into one of three specific aircraft training curricula: the maritime, jet, or helicopter pipelines. The ASTB was not validated to predict performance or attrition of SNAs beyond this 'pipelining.' The student NFO curriculum is identical to the student pilot training during ground school at NASC. Upon graduation from NASC, student NFOs are sent to squadron VT-10 at NAS Pensacola, Florida for their basic training. This training includes approximately 10 h of familiarization/stick time in the T-34C Mentor, specific coursework on systems and procedures for the T-39 Sabreliner, and a several basic navigation flights in the T-39. This concludes their training at VT-10. Subsequent to this phase, student NFOs are pipelined to the jet, strike, or maritime curricula. The ASTB was only validated to predict performance and attrition of student NFOs through their VT-10 training.

The current form of the ASTB test is the 1992 revision of the paper-and-pencil selection tests used by the U.S. Navy since World War II. The two equivalent forms of the ASTB take approximately 1 h and 45 min to administer and include six timed subtests. The subtests include the Math/Verbal Test (MVT), Mechanical Comprehension Test (MCT), Spatial Apperception Test (SAT), Aviation and Nautical Information Test (ANI), Biographical Inventory Test (BI), and the Aviation Interest Test (AI). The MVT contains 37 questions and evaluates basic math skills and paragraph comprehension. All candidates are given 35 min to complete this section. The MCT evaluates the candidates' knowledge of basic mechanical principles and physics and consists of 30 questions, with 15 min allotted for completion. The SAT measures a candidate's ability to match an "inside-out" or "cockpit view" to an outside-in" or "wingman's view." Candidates are shown an aerial view of coastal terrain, which represents the view from the cockpit. They are then shown a series of pictures with aircraft in different attitudes and must determine which attitude matches the pictured terrain. This subtest consists of 35 questions and must be completed in 10 min. The ANI test consists of 30 questions to be answered in a 15 min time limit. The ANI tests a candidate's specific knowledge about aviation and nautically-related material. The BI is a standard biographical inventory in which candidates respond to questions regarding their academic background and performance, as well as their life experiences. Candidates are given 20 min to complete this portion of the test. The Aviation Interest subtest (AI) has not been validated against flight performance but was included in the test battery to allow for the continued validation of the alternate test items that are included in this section. The goal was to use these items in future versions of the ASTB.

For applicants in Naval and Marine Corps aviation programs, the subtest scores on the ASTB are combined into a total of six composite scores: Academic Qualification Rating (AQR), Pilot Flight Aptitude Rating (PFAR), Flight Officer Flight Aptitude Rating (FOFAR), Pilot Biographical Inventory (PBI), Flight Officer Biographical Inventory (FOBI), and Officer Aptitude Rating (OAR). The AQR, PFAR, and FOFAR are weighted composites of the MVT, MCT, SAT, and ANI. The composite scores are converted from their raw form and reported as a stanine score. The original validation sample was used for the creation of the current stanine distribution. The AQR was validated to predict academic performance in aviation preflight indoctrination and ground school. The PFAR is validated against a criterion of primary pilot performance (i.e., flight and simulator performance), and the FOFAR with a criterion of basic NFO flight and simulator performance. The OAR is a composite score including only the MVT and MCT tests. The OAR was not validated for the aviation community and is not used for selection in the pilot or NFO programs. The PBI and FOBI are derived separately from the BI subtest. The PBI was validated to predict student pilot attrition from primary training, and the FOBI was validated to predict student NFO attrition from basic training. The predictive validity of the ASTB subtests for flight performance and success in training range from .23 to .40 (Multiple R), depending on the subtest and specific criterion used for analysis (Frank & Baisden, 1992). For the present study, raw scores for all subtests are used as the dependent measures.

The Computerized ASTB (C_ASTB)

The C_ASTB is a Windows-based/C++ program developed over a 3-year period at the NAMRL. The goal of the project was to duplicate, as closely as possible, the paper-and-pencil version of the existing ASTB for eventual transition to the operational community. The system runs on an IBM- or PC-compatible computer (486/33 MHZ or greater) and requires a VGA monitor, a standard keyboard, and a mouse. The entire test is menu-driven, and responses can be made via keyboard or mouse. For the purposes of this study, all responses were made using a single click with the left mouse. Each subtest begins with an instructions and examples page. At the end of the instructions page, participants are warned that the test section will begin when they move from that page. When participants choose to move on from the instructions page, the timer for that section is begun. Each test item fits on a single screen, so no scrolling is necessary. The test item images were all created to mirror those of the paper-and-pencil test. Additional control features for each item/screen are located at the bottom of each screen, and present the participant with three options. They may either click on a button bar to (1) move to the next test item; (2) move back to the previous item; or, (3) mark the item for later review. An example of the screen format can be seen in Fig. 1. Participants are given the same time limits on the computerized subtest as they are allotted in the paper-and-pencil versions of the subtests. If they fail to complete the test section within the allotted time, the section will automatically close out and the next test section will begin. If they complete the test section prior to the end of the allotted time, participants are informed as to the time remaining for that particular subtest and the number of unanswered questions. They are prompted to either review all of the test questions within that section, review only the marked items within the section, or to exit the current subtest. Once participants exit, they cannot return to any previous subtest. If they make no response, the subtest will close out when the timer runs out.

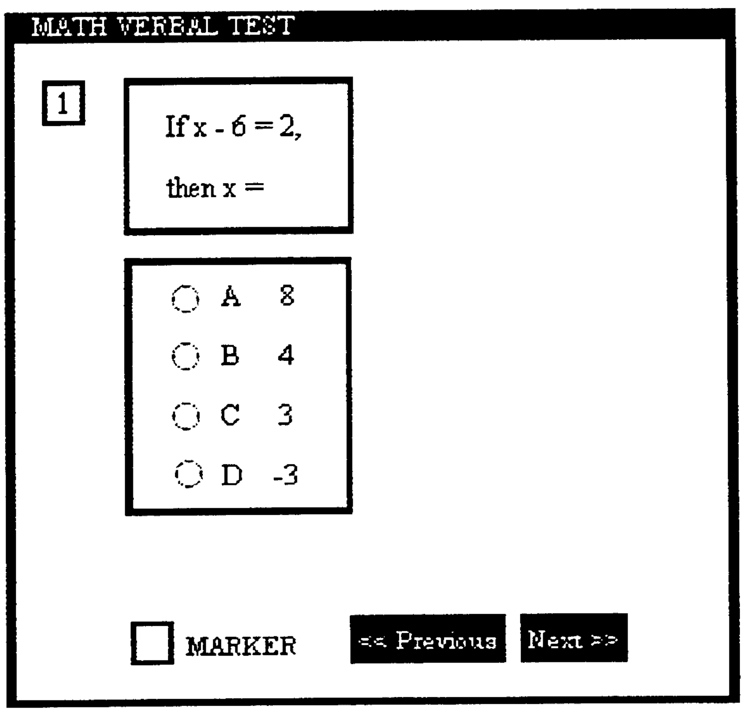


Figure 1. Example of test screen format.

The ASTB - Alternate (ALT)

To evaluate the influence of computerization and the specific features of this system (i.e., graphics quality, program interface features, etc.) independent of experience effects, a test battery which is novel to all participants, was needed. Therefore, two forms of a paper-and-pencil alternate test battery, the ALT, were developed in-house. The ALT was developed specifically to measure similar underlying constructs as the ASTB, and, therefore, similar subtests were created for the ALT forms. The ALT, MCT, and MVT subtests were created by NAMRL personnel psychologists and consisted of simple arithmetic, sentence comprehension, and mechanical/physical principles questions, respectively. The ALT, BI, and ANI were modeled after their corresponding ASTB subtests, and were likewise created by NAMRL psychologists. The SAT portion of the ALT, however, included test items that were different conceptually from the ASTB SAT. The ALT SAT consisted of folding figures, figure rotation, and Raven's Matrices types of questions. The number of test items in each section and the time to complete each ALT subtest was identical to the corresponding ASTB section. Both paper-and-pencil and computerized versions of this test were created and used in this study.

DESIGN

A split plot design (see Table 1 below) was used with three experimental treatments labeled as A (TEST), B (FORM), and C (MEDIUM), with two levels to each treatment. Treatment A was the test instrument with levels a_1 (ASTB) and a_2 (ALT). Treatment B was the different forms of each test instrument where b_1 was Form I and b_2 was Form II. Treatment C was the different presentation medium where level c_1 was the paper-and-pencil format and c_2 was the computer-based test. Both B and C were between-participants' blocks, and A was a within-blocks treatment. Therefore, this design represented repeated measures for the experimental treatment A. Each set of subjects (S1 through S4) took both the ASTB and the Alternate version, but each set of participants had only one form of the test, which was presented through only one medium. Subjects reported at 0800 on the first day of testing for briefing and informed-consent procedures, and testing began at approximately 0830. Instructions for the computer-based and paper-and-pencil test were identical, except that instructions in the paper-and-pencil conditions were provided verbally by the experimenter. The procedure for administering the ASTB are delineated in the examiner's manual that comes with the test battery. These instructions were followed verbatim for the ASTB conditions. Computerized instructions were almost identical to those of the paper-and-pencil test, except where obvious modifications were needed. For example, explanations of how to maneuver through the screens with a mouse input, how to review a section, etc. were included in the C_ASTB. Similar instructions were used for the ALT paper-and-pencil and computerized forms.

Table 1. Experimental Design.

	Test		
Form x Medium	a_1	a_2	n
b_1c_1	S_1	S_1	20
b_1c_2	S_2	S_2	20
b_2c_1	S_3	S_3	22
b_2c_2	S_4	S_4	20

During the briefing procedures, the participants' previous ASTB test (PREV ASTB) information was obtained from the Naval Operational Medicine Institute. This information included both their previous test scores, which were used for their accession into the student naval aviation training program, as well as which form of the test they had previously received. Their previous test form determined which form of the ASTB they would receive in the current study. If they had previously taken Form I, on their first test day in the laboratory they would take Form II of the ASTB and the ALT (either paper-and-pencil or computerized versions, depending on their group assignment). Likewise, if they had previously taken Form II, they would take Form I. The order of presentation for the ASTB and ALT tests was counterbalanced between participants.

Participants were required to return the next day at 0830 for their second test session, which included administration of the test (ALT or ASTB) that they had not taken the previous day. The number of correct responses for each subtest was used as the dependent measures. Analysis of variance (ANOVA) and Pearson correlations were calculated using SPSS 6.0 for Windows. Results were considered significant for $p < .05$. No analyses were done on the BI data, because there were no 'correct' or 'incorrect' responses for these test items on the ALT form. In addition, participants' subtest scores from their original ASTB administration (PREV ASTB) were used in the analysis.

RESULTS

Descriptive Statistics

Eighty-one of the 82 participants had taken the ASTB prior to reporting to the NASC, and all but one of the subjects had received a 4-year college degree prior to arriving for training. This one individual had received his commission through an enlisted commissioning program. The mean scores and standard deviations on each of the four subtests for the three versions of the test can be seen in Table 2. The ASTB and ALT tests were administered during the course of the experiment so these mean scores were calculated over participants who took both Form I and Form II and both the paper-and-pencil and computerized formats of the tests. These tests were taken under fairly controlled laboratory conditions. The test labeled PREV was taken prior to the experiment (for example, at a recruiting station) as a standard paper and pencil test but otherwise under unknown controlled conditions.

Table 2. Means and Standard Deviations.

Test		Mean	Standard deviation	*N*
MVT	PREV	27.80	4.89	81
	ASTB	27.33	5.55	82
	ALT	23.63	5.27	82
MCT	PREV	22.22	3.82	81
	ASTB	21.88	4.00	82
	ALT	16.01	3.99	82
ANI	PREV	18.72	3.54	81
	ASTB	19.55	3.97	82
	ALT	21.33	3.36	82
SAT	PREV	26.94	5.43	81
	ASTB	26.65	5.69	82
	ALT	19.77	4.06	82

Analysis of Mean Scores

Differences in means due to treatment effects are identified through an ANOVA. The following results are presented in the form of standard ANOVA tables for the repeated measures design used in this study (Kirk, 1968). Significant F values in the ANOVA tables indicate statistically significant differences for mean test scores over the levels of the particular treatment effect. For significant interaction effects, a graph plotting the means involved in the interaction will be employed. The analysis for the overall score is given first, followed by separate analyses for each of the four subtests making up the overall score. For this experimental design, there are three main effects, FORM, MEDIUM, and TEST, three double interactions, FORM x MEDIUM, FORM x TEST, and MEDIUM x TEST, and one triple interaction, FORM x MEDIUM x TEST. The mean assessed by the ANOVA from a triple interaction comes from a specific treatment level for each of the three treatment factors, and the averaging is done only over the number of participants in that specific treatment combination. For example, choosing a specific treatment level for each of the three treatment factors might result in Form I (FORM) of the paper-and-pencil version (MEDIUM) of the ASTB (TEST). There were 20 subjects in this condition so the mean score is constructed over these 20 participants. A double interaction will average over participants and *both* levels of the treatment factor not involved in the specified interaction. In a MEDIUM x TEST interaction, the mean score is calculated over all participants and over both levels of FORM. Finally, for a main effect mean score, the average will be taken over all participants and over both levels of the other two treatment factors.

The first result (Table 3) shows the impact of the three experimental treatments on the overall combined score. The combined score is determined by adding together the scores on four subtests, 1) MVT (Math-Verbal), 2) MCT (Mechanical Comprehension), 3) SAT (Spatial Apperception), and 4) ANI (Aviation and Nautical Interest).

Table 3. ANOVA of Combined Score.

Source	*SS*	*df*	*MS*	*F*	*P*(*F* or greater)
Between subject					
Form	132.78	1	132.78	.59	.446
Medium	6.59	1	6.59	.03	.865
Form x Medium	163.91	1	163.91	.72	.398
Error Term	17677.50	78	226.63		
Within subject					
Test	8775.01	1	8775.01	127.94	.000
Form x Test	254.19	1	254.19	3.71	.058
Medium x Test	173.35	1	173.35	2.53	.116
Form x Medium x Test	2.40	1	2.40	.03	.852
Error term	5349.61	78	68.58		

The top-half of the ANOVA table lists the between subject treatments (FORM and MEDIUM) and their interaction (FORM by MEDIUM) under the Source heading and the bottom half lists the within subject treatment (TEST) with all of its interactions. The succeeding columns give the sum-of-squares (SS), the degrees of freedom (df), the mean square (MS), the *F* ratio, (*F*) and, finally, in the last column the probability of obtaining this value of the *F* ratio or higher (P(*F* or greater)). The standard convention of declaring significance when this value is less than .05 will be employed.

There is no main effect due to using two different forms on overall test scores and, more importantly, no main effect due to presenting the test in a paper and pencil format as opposed to presentation over the computer. In the bottom half of the table there is one highly significant *F* main effect indicating that the mean overall test scores for the ALT test are much lower than the mean overall test scores for the ASTB test. There were no significant interactions for either the Between treatments or the Within treatments.

As a further confirmation of this analysis derived from the ANOVA tables, the means of the overall test scores are presented in the following two tables. Table 4 shows a two way table of FORM x TEST. The two parallel versions of the test captured in Form I and Form II do not differ significantly on either the ALT test or the ASTB test. Similarly, Table 5 presents the breakdown of means for MEDIUM x TEST, where again no significant difference was found between the overall scores for paper-and-pencil presentation and computer presentation. However, it can be observed from both tables that the difference in overall test scores between the ALT test and the ASTB test is relatively large. The relationship can also be seen graphically in Fig. 2.

Table 4. Means of Overall Scores for FORM by TEST.

	ALT Test	ASTB Test
Form I	82.95	95.10
Form II	78.66	95.79

Table 5. Means of Overall Scores for MEDIUM by TEST

	ALT Test	ASTB Test
Paper and Pencil	81.47	94.22
Computer	79.98	96.68

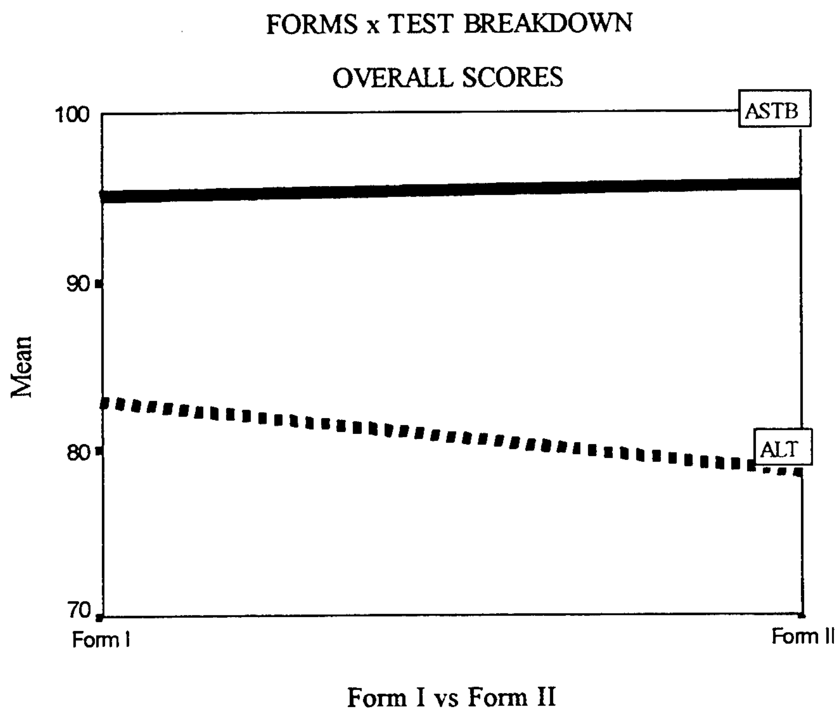


Figure 2. Means for the overall test scores when broken down over the two levels of the FORM treatment.

Likewise, when the means are presented as broken down over the MEDIUM treatment (see Fig. 3), we found there is no essential difference between scores obtained when the test was administered via the computer as compared to administration as a paper and pencil test. The two TEST levels differed by roughly 15 points with a higher score indicating better performance on the test.

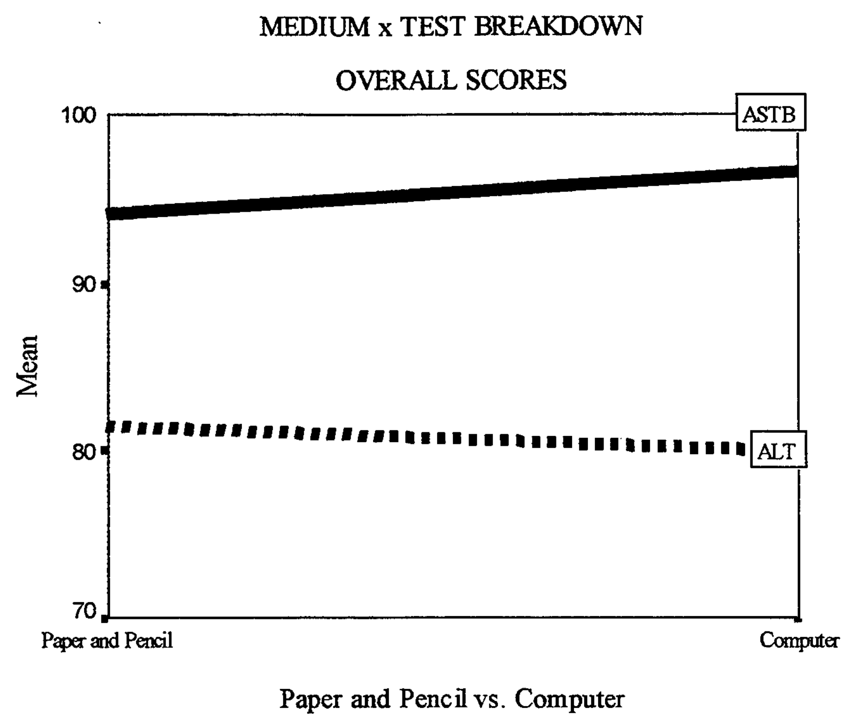


Figure 3. Means for the overall test scores when broken down over the two levels of the MEDIUM treatment.

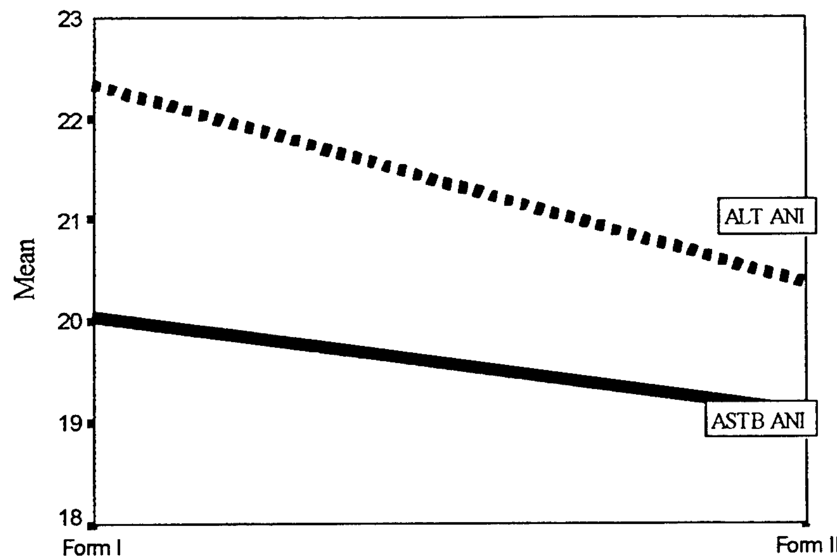
Results are not as clear cut as for the overall scores. Some significant differences in means due to the main effect of treatment conditions or due to the interactions of main effects. The following four tables are presented in order of increasing complexity of interpretation. First, the ANOVA table for the ANI subtest is given in Table 6. A difference in ANI scores due to the main effect of FORM with Form I having a mean of 21.19 versus a mean of 19.73 for Form II. The experimental treatment due to TEST (*i.e.*, whether the alternate version of the ASTB or the original ASTB was administered) will be significant in all of the analyses of the subtests just as it was for the overall score. For the ANI, however, in contrast to the overall score and the other subtests, the mean score was higher for the alternate version than for the ASTB itself.

Table 6. ANOVA of ANI subtest.

Source	*SS*	*df*	*MS*	*F*	*P*(*F* or greater)
Between subject					
Form	85.25	1	85.25	4.39	.039
Medium	14.40	1	14.40	.74	.392
Form x Medium	1.16	1	1.16	.06	.808
Error term	1513.13	78	19.40		
Within subject					
Test	132.61	1	132.61	18.51	.000
Form x Test	11.28	1	11.28	1.57	.213
Medium x Test	1.25	1	1.25	.17	.677
Form x Medium x Test	.41	1	.41	.06	.812
Error term	558.75	78	7.16		

DIFFERENCE FOR FORMS

ANI



Form I vs. Form II

Figure 4. Means for the two ANI test scores when broken down over the two levels of the FORM treatment.

Next in order of complexity in interpretation is the SAT subtest as shown in Table 7. Here, in addition to the difference in scores due to the TEST treatment effect (with the ALT SAT being lower than the ASTB SAT), there is an interaction of TEST with MEDIUM. The manner in which the tests were presented to the participants did make a difference in this case as can be seen in Fig. 5.

Table 7. ANOVA of SAT subtest.

Source	*SS*	*df*	*MS*	*F*	*P*(*F* or greater)
Between Subject					
Form	51.70	1	51.70	1.86	.177
Medium	2.58	1	2.58	.09	.761
Form x Medium	2.09	1	2.09	.08	.785
Error Term	2170.45	78	27.83		
Within Subject					
Test	1932.67	1	1932.67	94.86	.000
Form x Test	44.83	1	44.83	2.20	.142
Medium x Test	83.51	1	83.51	4.10	.046
Form x Medium x Test	17.11	1	17.11	.84	.362
Error Term	1589.17	78	20.37		

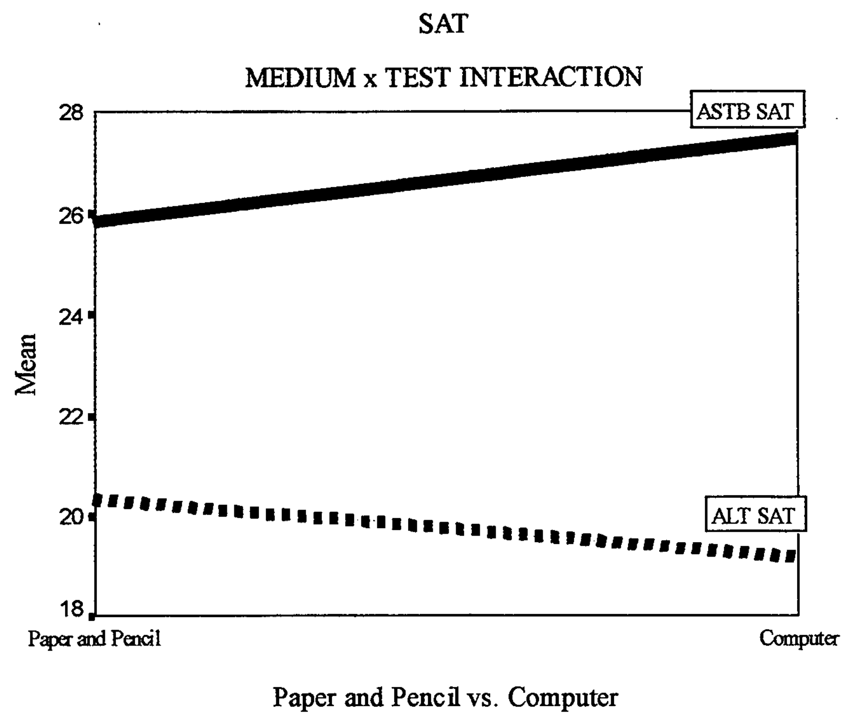


Figure 5. Means for the two SAT subtest scores when broken down over the two levels of the MEDIUM treatment.

The ANOVA for the MCT subtest is shown in Table 8. There were significant main effects for the FORM and TEST variables, as well as a significant FORM x TEST interaction.

Table 8. ANOVA of MCT subtest.

Source	*SS*	*df*	*MS*	*F*	*P*(*F* or greater)
Between Subject					
Form	93.21	1	93.21	4.47	.038
Medium	2.89	1	2.89	.14	.711
Form x Medium	28.48	1	28.48	1.37	.246
Error Term	1627.05	78	20.86		
Within Subject					
Test	1377.97	1	1377.97	150.11	.000
Form x Test	80.48	1	80.48	8.77	.004
Medium x Test	7.47	1	7.47	.81	.370
Form x Medium x Test	24.73	1	24.73	2.69	.105
Error Term	716.01	78	9.18		

The graphical display of the FORM x TEST interaction (see Fig. 6) indicates that scores are level across the two forms for the ASTB MCT while they decline from Form I to Form II for the ALT MCT. The main effect due to the two different tests is clearly visible. The ALT MCT, like the other subtests except for the ANI subtest, is a harder test for which the subjects' scores are lower.

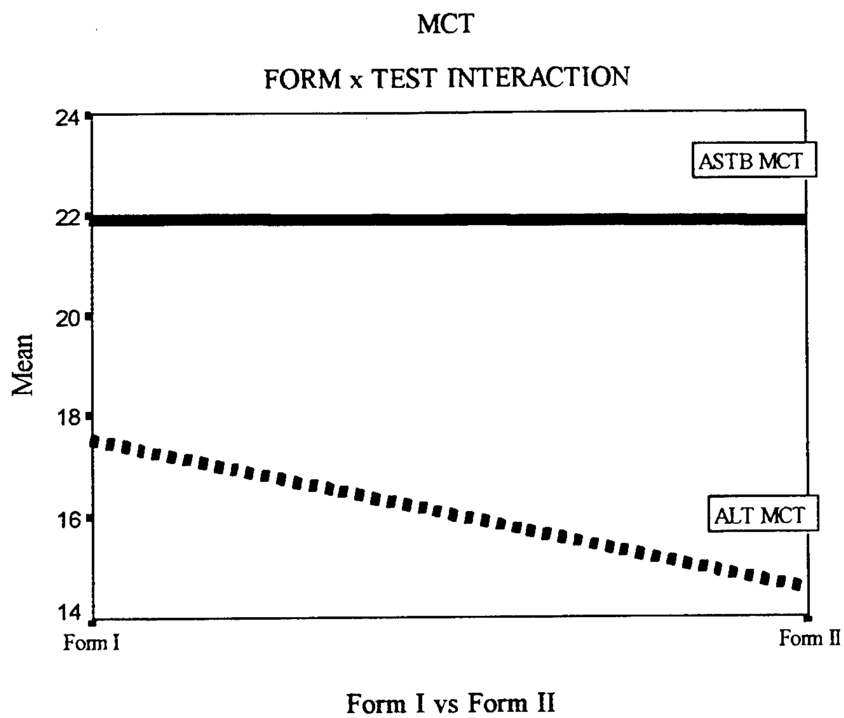


Figure 6. Means for the two MCT subtest scores when broken down over the two levels of the FORM treatment.

ANOVA results for the MVT are given in Table 9. There were no significant main effects due to FORM or MEDIUM. However, both a significant double interaction of MEDIUM x TEST and a significant triple interaction of FORM x MEDIUM x TEST.

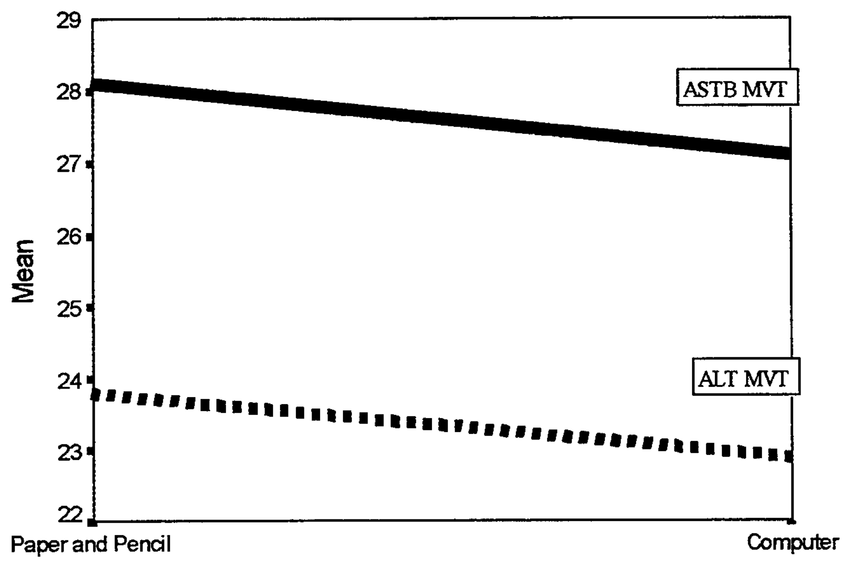
Table 9. ANOVA of MVT subtest.

Source	*SS*	*df*	*MS*	*F*	*P*(*F* or greater)
Between Subject					
Form	.03	1	.03	.00	.979
Medium	1.29	1	1.29	.03	.864
Form x Medium	24.44	1	24.44	.56	.456
Error Term	3401.86	78	43.61		
Within Subject					
Test	581.18	1	581.18	38.64	.000
Form x Test	9.50	1	9.50	.63	.429
Medium x Test	62.11	1	62.11	4.13	.046
Form x Medium x Test	67.25	1	67.25	4.47	.038
Error Term	1173.23	78	15.04		

Fig. 7 shows the MEDIUM x TEST double interaction for FORM I, and Fig. 8 shows the double interaction for FORM II. A comparison of the two graphs illustrates the triple interaction. It is clear from these two figures that FORM I exhibits a parallel reaction to the paper-and-pencil mode of administration versus the computer mode of administration. The scores decline when the test is given by computer for both the ASTB MVT and the ALT MVT. For Form II, the ASTB MVT test reacts differently to computer presentation than does the ALT MVT. The scores increase for the ASTB MVT in the computer medium over the paper-and-pencil medium while the scores decrease for the ALT MVT.

FORM x MEDIUM x TEST INTERACTION

FORM I MVT

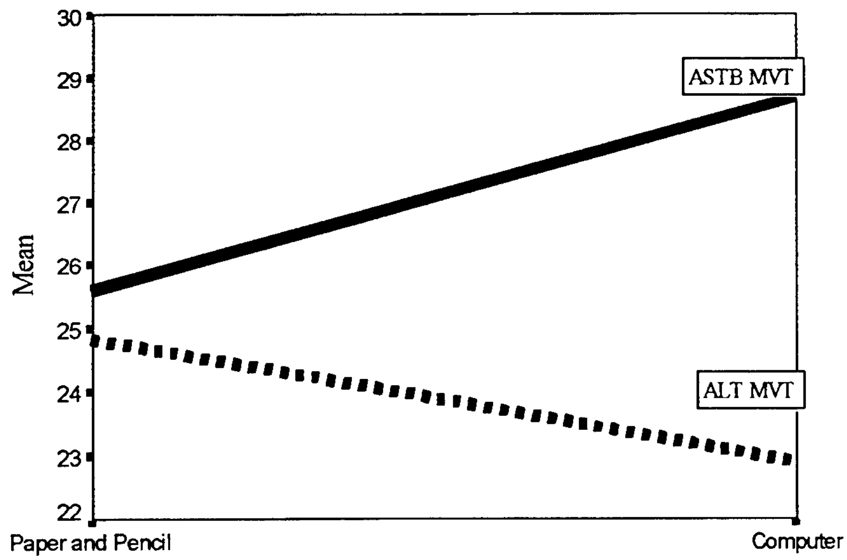


Paper and Pencil vs. Computer

Figure 7. Means for the two MVT subtest scores when broken down over the two levels of the MEDIUM treatment. The means are further subdivided by the FORM treatment with this graph showing Form I.

FORM x MEDIUM x TEST INTERACTION

FORM II MVT



Paper and Pencil vs. Computer

Figure 8. Means for the two MVT subtest scores when broken down over the two levels of the MEDIUM treatment. The means are further subdivided by the FORM treatment with this graph showing Form II.

In summary, the important treatment effect for this study was presentation medium; that is, did the manner in which the test was administered (in the traditional paper-and-pencil version or in the computerized format) significantly affect the mean scores? For the overall test scores, the answer is no. When scores for the individual subtests are analyzed separately, presentation medium sometimes does have an impact. Specifically, it had an impact on two of the four subtests, the SAT and the MVT. Giving the test in a computerized format had a positive impact on scores for the ASTB SAT and a negative impact on the ALT SAT. The same was true for Form II of the MVT, whereas for Form I of the MVT, the computerized format depressed scores for both versions of the test. However, MEDIUM was never a significant main effect in any of the analyses so it may be that presenting test material via the computer will have an impact (positive or negative) in only certain circumscribed conditions. Further research will attempt to determine exactly what those special conditions might be.

CORRELATIONAL STRUCTURE

Correlations were computed between each of the subtest scores on the ASTB given at NAMRL, the ALT test scores, and the original ASTB that participants had taken for their initial selection to the student naval aviator program. The latter test scores were received from the participants' accession points (i.e., the U.S. Naval Academy, Navy ROTC programs, or Navy Recruiting District Offices). As can seen in Table 10, there were significant correlations between the ALT MVT, MCT and ANI and their corresponding subtest of the ASTB administered at NAMRL. The ALT SAT was significantly correlated with the ASTB MCT and MVT administered at NAMRL, but not the corresponding ASTB SAT. For the PREV ASTB subtests, the PREV MVT scores were significantly correlated with performance on the ASTB MVT. No other correlations between this test and the ASTB and ALT administered at NAMRL were evident. Interestingly, the pattern of within-test correlations observed were very different for the three sets of test scores (see Tables 11-13). The ASTB and ALT showed a similar pattern, with significant correlations between the MVT, MCT, and SAT for the ALT test, and between the

ANI, MCT and MVT for the ASTB. The ASTB MCT was also significantly correlated with the SAT. The PREV ASTB also showed significant correlations between the MCT and the ANI and MVT test scores, but not between the SAT and any other subtest.

Table 10. Correlation Coefficients.

	PREV				ALT			
ASTB	**ANI**	**MCT**	**MVT**	**SAT**	**ANI**	**MCT**	**MVT**	**SAT**
ANI	-.0097	.1368	.0433	.0076	.4840**	.0714	.1171	.0065
MCT		.1251	.0992	.1689		.3559**	.3261*	.3730**
MVT			.3785*	.0563			.4452**	.3977**
SAT				.1891				.1322
PREV								
ANI					-.0618	.0054	.0938	.1049
MCT						.0771	.1646	.1066
MVT							.1837	.0656
SAT								.0643

* $p < .01$; ** $p < .001$

Table 11. Correlation Coefficients for the ALT Test.

ALT	ANI	MCT	MVT	SAT
ANI		.1021	-.0092	-.0660
MCT			.4164**	.4438**
MVT				.4413**
SAT				

** $p < .001$

Table 12. Correlation Coefficients for the ASTB.

ASTB	ANI	MCT	MVT	SAT
ANI		.4757**	.2758*	.1092
MCT			.5489**	.2938*
MVT				.0893
SAT				

* $p < .01$; ** $p < .001$

Table 13. Correlation Coefficients for the PREV ASTB.

PREV	ANI	MCT	MVT	SAT
ANI		.4021**	.0400	.1049
MCT			.3994**	.0447
MVT				-.1374
SAT				

** $p < .001$

DISCUSSION

A definitive answer on the equivalence of the predictive validity of the computerized ASTB and the paper-and-pencil ASTB will not be available until participant training performance data are collected. The current study, however, does suggest that the chosen Windows-based format is compatible with the Navy's current ASTB, without the need for score transformations. Although some marginally significant interactions of medium and test (ASTB versus ALT) were observed, we found no consistent differences (positive or negative) in subtest performance when comparing the computerized and paper-and-pencil ASTB or ALT tests. This was true even on the SAT subtests, which consist exclusively of graphics and are presumably more likely candidates for degraded performance. Even so, the SAT subtest images are not exceedingly complex, and it is possible that more detailed images would lead to performance decrements when displayed on a standard monitor. Some other reasons for the observed equivalence in this study include the specific design features of the system (i.e., marking, backtracking, etc.), and the procedural restriction to a single input device. There is some debate on whether previous computer experience or fear of computers significantly impacts performance on computer-based tests that do not involve a strong psychomotor component (Hofer & Green, 1985; Wise & Plake, 1989; Federico, 1992). In the present study, participants reported a wide range of previous computer experience, but this did not seem to be a factor in performance on the tests. Interestingly, even though some participants initially reported no computer experience, further questioning revealed that only two participants had never used a word processing program or a mouse prior to entering the study. For the operational ASTB, a similar 'real world' population is expected, since applicants have a minimum of 2-3 years of college education prior to taking the test.

When comparing the ALT test to the ASTB, as averaged over both presentation medium and form, the in-house test was obviously more difficult than the ASTB. This was true for all of the subtests, with the exception of the ANI. In contrast, no differences were seen between the PREV ASTB and the ASTB taken in the laboratory. This suggests that the lower performance on the ALT was due to the greater difficulty of the test items and not duc to lower motivation for participants to perform well on the laboratory-administered tests. Another explanation for these results is that the equivalent performance on the PREV ASTB and laboratory ASTB have independent causes. For example, performance on the laboratory ASTB may be influenced by previous ASTB test exposure, and the PREV ASTB performance may be elevated due to motivational pressures, studying prior to testing, and/or coaching.

The unique error associated with all three versions of the test was quite high, or what amounts to the same thing, the test reliabilities were quite low. This makes the whole question of determining psychometric equivalency that much more difficult. Because the correlations among the subtests for the previously administered ASTB and the laboratory administered ASTB were generally not significant (see left half of Table 10), it is hard to assess when controlled experimental factors like presentation medium are going to have an impact on mean performance. Putting it succinctly, if two administrations of ostensibly the same test vary so greatly, then how can any experimental manipulation be shown to have an effect on psychometric equivalency? If one is forced by definition, to accept that the two forms of the ASTB are psychometrically equivalent, then any experimental manipulation, no matter how disruptive, would be considered benign in terms of the correlational structure. Presenting test items over the computer as opposed to paper-and-pencil is not going to add any greater error to scores than already presented giving the same test by the same medium at a later date.

The correlations within each test showed different relationships between the subtests of the ASTB and PREV ASTB. This suggests that the factor structure of the ASTB differed for the field and laboratory administrations. When the database sample size increases sufficiently, structural equation modeling (including a confirmatory factor analysis) will be done and the predictive validation of the computer-based ASTB will be completed.

REFERENCES

American Psychological Association Committee on Professional Standards and Committee on Psychological Tests and Assessment. (1986). Guidelines for Computer-Based Tests and Interpretations. Washington, D.C.: Author

Federico, P.A. (1992). Assessing Semantic Knowledge Using Computer-Based and Paper-Based Media. Computers in Human Behavior, 8, pp. 169-181.

Frank, L. & Baisden, A. (1994). The 1992 Navy and Marine Corps Aviator Selection Test Battery Development. Presented at the 1994 Annual Meeting of the Military Testing Association, Williamsburg, VA.

Green, B.F. (1991). Guidelines for Computer Testing. The Computer and the Decision-Making Process. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 245-273.

Harrell, T.H., Honaker, L.M., Hetu, M., & Oberwager, J. (1987). Computerized Versus Traditional Administration of the Multidimensional Aptitude Battery-Verbal Scale: An Examination of Reliability and Validity. Computers in Human Behavior, 3, pp. 129-137.

Hofer, P.J. & Green, B.F. (1985). The Challenge of Competence and Creativity in Computerized Psychological Testing. Journal of Consulting and Clinical Psychology, 53, pp. 826-838.

Kiely, G.L., Zara, A.R. & Weiss, D.J. (1986). Equivalence of Computer and Paper-and-Pencil Armed Services Vocational Aptitude Battery Tests. Air Force Human Resources Laboratory Final Technical Paper (AFHRL-TP-86-13), Brooks Air Force Base, TX.

King, W. C. & Miles, E. W. (1995). A Quasi-Experimental Assessment of the Effect of Computerizing Noncognitive Paper-and-Pencil Measurements: A Test of Measurement Equivalence. Journal of Applied Psychology, 80, pp. 643-651.

Mead, A.D., & Drasgow, F. (1993). Equivalence of Computerized and Paper-and-Pencil Cognitive Ability Tests: A Meta-Analysis. Psychological Bulletin, 114, 3, pp. 449-458.

Moreno, K.E., Wetzel, C.D., McBride, J.R., & Weiss, D.J. (1984). Relationship Between Corresponding Armed Services Vocational Aptitude Battery (ASVAB) and Computerized Adaptive Testing (CAT) Subtests. Applied Psychological Measurement, 8, 2, pp. 155-163.

Van de Vijver, F.J.R. & Harsveld, M. (1994). The Incomplete Equivalence of the Paper-and-Pencil and Computerized Versions of the General Aptitude Test Battery. Journal of Applied Psychology, 79, 6, pp. 852-859.

Vansickle, T.R., Kimmel, C., & Kapes, J.T. (1989). Test-Retest Equivalency of the Computer-Based and Paper-Based Versions of the Strong-Campbell Interest Inventory. Measurement and Evaluation in Counseling and Development, 22, pp. 88-93.

Ward, W.C. (1984). Using Microcomputers to Administer Tests. Educational Measurement: Issues and Practices, summer, pp. 16-20.

Wise, S.L. & Plake, B.S. (1989). Research on the Effects of Administering Tests via Computers. Educational Measurement: Issues and Practice, 8, 3, pp. 5-10.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	4 March 1998	

4. TITLE AND SUBTITLE
The Effect of Presentation Medium on Pilot Selection Test Battery Scores

5. FUNDING NUMBERS
62233N
MM33P30-017602

6. AUTHOR(S)
S. Biggerstaff, C.A. Portman, D.J. Blower, and A. Chapman

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Aerospace Medical Research Laboratory
51 Hovey Road
Pensacola Fl 32508-1046

8. PERFORMING ORGANIZATION REPORT NUMBER
NAMRL-1400

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Naval Medical Research and Development Command
National Naval Medical Center
Building 1, Tower 12
8901 Wisconsin Avenue
Bethesda, MD 20889-5606

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The use of computer-based testing systems for selection and classification has gained popularity in both the civilian and military world. However, there are several issues related to the transition from paper-and-pencil tests to computerized testing. The American Psychological Association (APA) has set guidelines to be followed to ensure both qualitative and quantitative equivalence of new test formats prior to their use in applied settings. Attention must be paid to the psychometric properties of the test design and the human-machine interface to produce a reliable and valid test. Other peripheral issues such as visual display qualities and input devices must also be carefully considered. The purpose of the study was to conduct an initial evaluation of the equivalence between the current paper-and-pencil version of the U.S. Navy's Aviation Selection Test Battery (ASTB) and a Windows-based version of the ASTB. Eighty-two U.S. Navy aviation flight school candidates at the Naval Aviation School's Command (NASC) volunteered to participate. Participants were given either the current paper-and-pencil version or the computer-based version of the ASTB or an alternate test developed to measure similar psychological attributes. The results showed that there were no main effects due to medium (paper-and-pencil vs. computer-based), although performance on the alternate test was significantly worse than on the ASTB. Predictive validation of the computerized ASTB will not be completed until participant training performance data is collected.

14. SUBJECT TERMS
Aviation selection, Selection test battery, ASTB, Pilot selection, Computer-based testing, Presentation

15. NUMBER OF PAGES
19

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102